Water Quality Diagnosis Using Classification

Data 1030 Final Report

Jin Hyeok Noh

Brown University

jin_hyeok_noh@brown.edu

GitHub link: https://github.com/jinnoh47/data1030-project/tree/main

1. Introduction

Getting access to safe drinking water is vital to humans' health, a fundamental human proper and a factor of adequate coverage for health protection. That is critical as a health and improvement trouble at a countrywide, local and nearby stage. In some areas, it has been shown that investments in water supply and sanitation can yield a net economic advantage since the reductions in adverse fitness effects and health care fees outweigh the charges of assigning the interventions.

The dataset was provided in Data Portal by the Republic of South Korea. There are 2036 data points and 43 features in this data set. The target variable is potability data indicates that zero is potable and one means not potable. The goal is to examine the water supply in various locations, identify whether water is potable or not, determine which factor impacts water potability, and create a machine learning development model for prediction. It is a classification problem, so utilizing various classification models such as Random Forest Classifier, XGBoost, Logistic Regression, and Support Vector Classifier will improve the prediction accuracy. Unfortunately, there was no public or publications about the data that has been used, but there was a research paper called "*A Study on the Turbidity Estimation Model Using Data Mining Techniques in the Water Supply System."* Giving hints about turbidity might be an essential factor.

2. Exploratory Data Analysis (EDA)

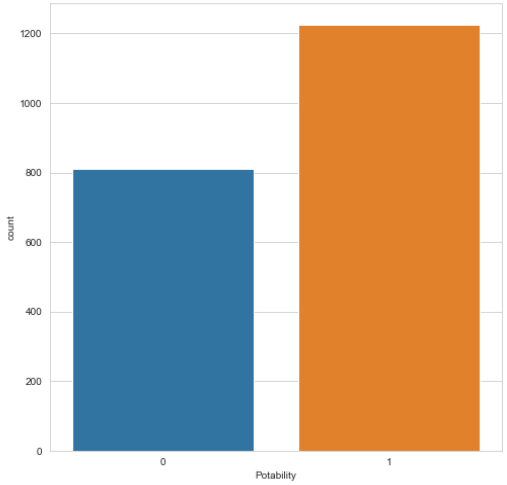

Figure 1. Frequency plot

Figure one displays the bar plot of the target variable. Bar plot is identical for using these types of situations. For example, the target variable is potability data indicating that zero is potable, and one means not potable even if the data type was an integer. Therefore, it should consider categorical data.

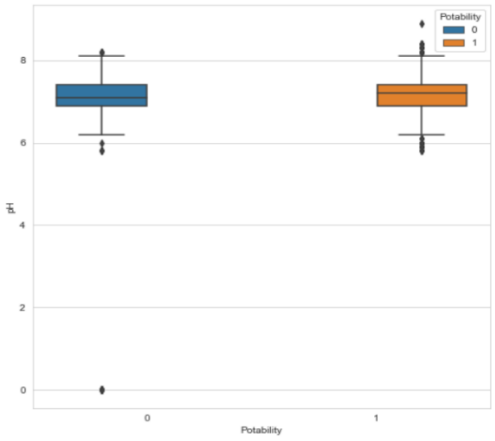

Figure 2. Box plot Potability and pH

This figure displays the box distribution of the ph and potability. Data points display some outliers, where water is potable or not potable, but despite human consumable water or not. There is not much big of a difference between the range of ph. This plot implies that other features involve water potability.

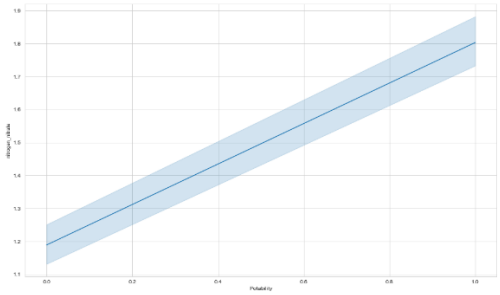

Figure 3. Line plot

This figure displays the line plot between potability and nitrogen nitrate. The plot indicates that nitrogen nitrate increases when water is not drinkable. This plot shows that there is a correlation between potability and nitrogen nitrate.

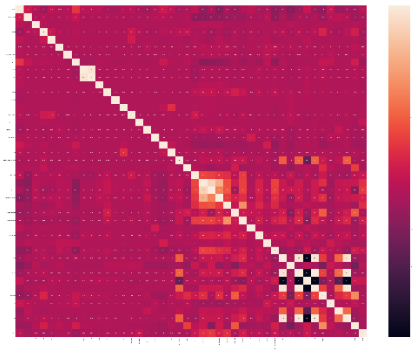

Figure 4. Heatmap correlation plot

This figure displays a correlation between columns using a heatmap. The heatmap shows that most of the correlations between columns are low, which means they are independent do not have linear regression to other features. Darker color indicates that it is negatively correlated with the brighter color, which is likely correlated.

3. Method

3-1 Data Preprocessing, Data Splitting

It is likely to understand that this data is in i.i.d form by reflecting on the data collection source. Within the data preprocessing stage, the splitting step allocated 20% of the dataset for testing, and the other 80% of the dataset was split into five folds. This means KFoldsSplitting(n=5).One is used for validation in every instance of cross-validation, and the other four folds are used for training purposes. The preprocessor is implemented as StandardScaler for every feature because continuous exploratory data analysis shows that all features are not reasonably bounded. It means that MinMaxEncoder is not suitable for this dataset preprocessing. As a result, forty-three features are preprocessed, and the target variable label is encoded into two categories.

After prepossessing progress data to implement different machine learning models, since the problem is about the classification analysis computing the baseline accuracy score for the prediction is reasonable. As mentioned in figure one, the data set is balanced roughly enough about a four-to-six ratio. Getting correct predictions for two classes is vital for this project, which is why the accuracy metric was selected.

3-2 Model Prediction

For the baseline accuracy rate, the portion of the larger class between the two is equal to 60.21%, which is manageable still considering the dataset is relatively small choosing four methodologies and fitting them with a random training dataset to decide which model would be further researched. This is a result of machine learning model predictions without parameter tuning.

Machine Learning algorithms	Logistic Regression	Random Forest	Support Vector Classifier	XGBoost
Accuracy Score	0.713235	0.884804	0.600490	0.889706

Table 1. result of machine learning model predictions without hyperparameter tuning

4. Results

This section displays which methodologies and the hyperparameter tunned showed improvement above baseline and accuracy score.

4-1 Logistic Regression

Test Baseline:	0.6200980392156863
Test Score:	0. 0.7769607843137255
Standard Deviation of Accuracy	+/- 0.01
Optimal Parameter:	'logisticregression__penalty': ['l1','l2'],'logisticregression__C': np.logspace(-4, 4, 20) ,solver='saga',max_iter=10000

Table 2. Logistic Regression Hyperparameter

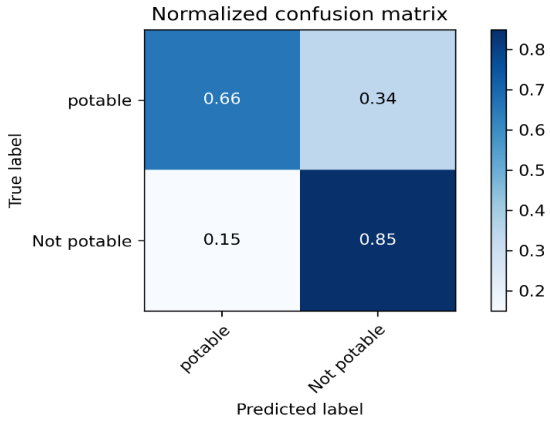

Figure 5. Normalized confusion matrix for Logistic Regression

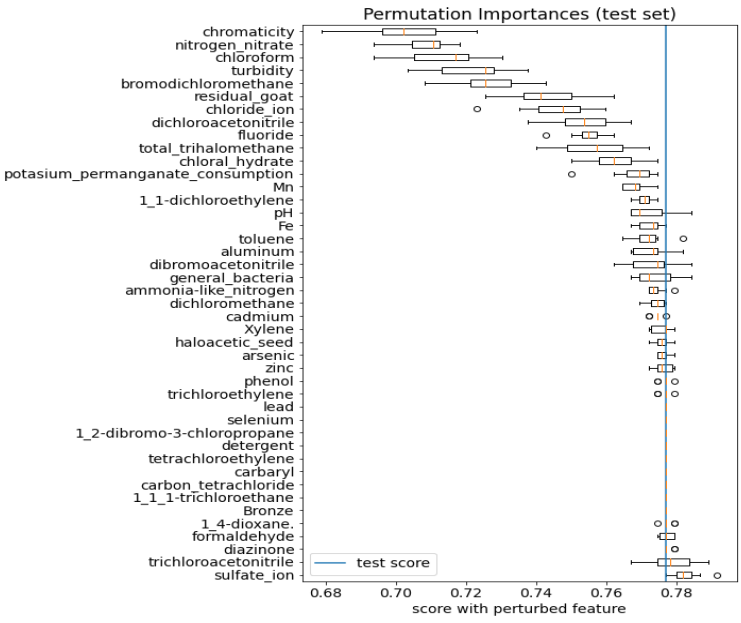

Figure 6. Feature importance for Logistic Regression

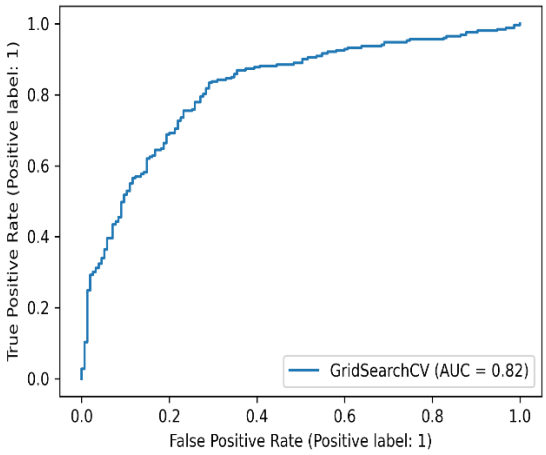

Figure 7. ROC curve for Logistic Regression

Using the parameters, above the baseline accuracy of Logistic Regression has been increased by around 2%. Feature importance shows that sulfate ion is the most crucial feature. Also, by plotting the ROC curve, AUC can reach 0.82.

4-2 Random Forest Classifier

Test Baseline:	0.6200980392156863
Test Score:	0.8700980392156863
Standard Deviation of Accuracy:	+/- 0.02
Optimal Parameter:	'randomforestclassifier__max_depth': [1, 3, 10, 25, 50], 'randomforestclassifier__max_features': [0.5,0.75,1.0]

Table 3. Random Forest Classifier Hyperparameter

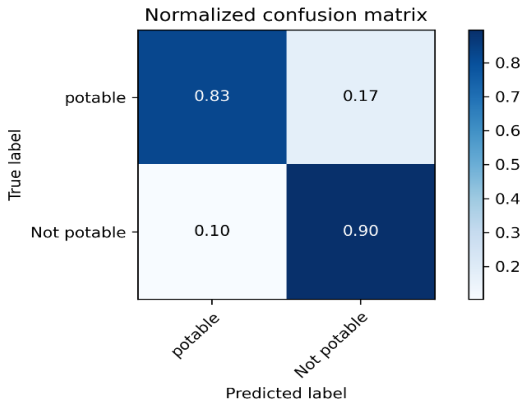

Figure 8. Normalized confusion matrix for Random Forest Classifier

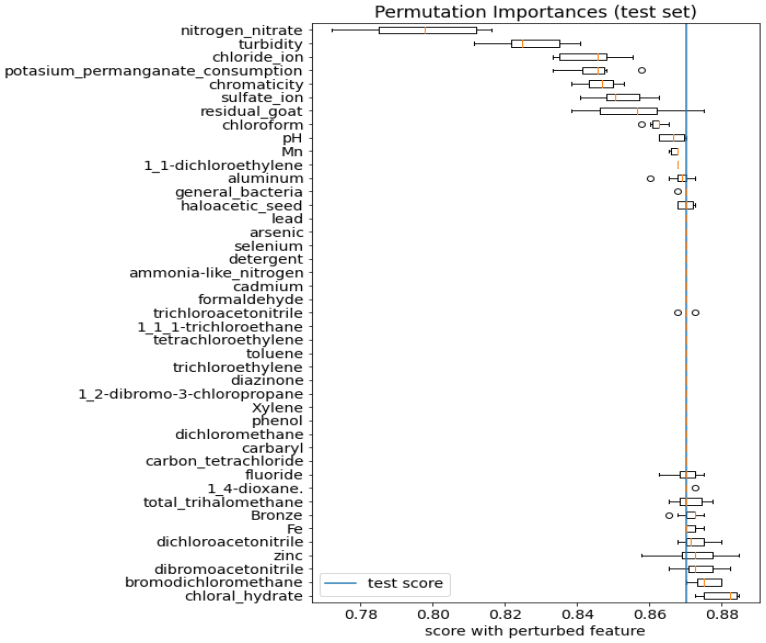

Figure 9. Feature importance for Random Forest Classifier

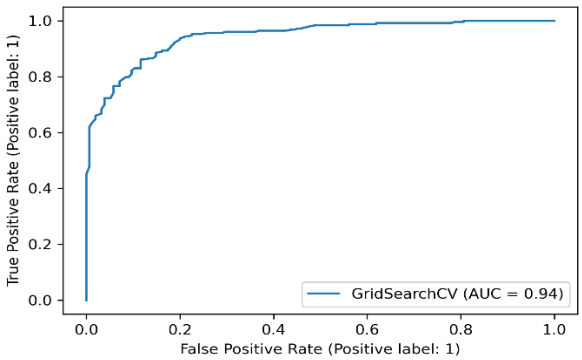

Figure 10. ROC curve for Random Forest Classifier

Using the parameters, above the baseline accuracy of the Random Forest classifier has been increased by around 2%. Feature importance shows that chloral hydrate is the most crucial feature. Also, by plotting the ROC curve, AUC can reach 0.94.

4-3 Support Vector Classifier

Test Baseline:	0.6200980392156863
Test Score:	0.8700980392156863

Standard Deviation of Accuracy:	+/- 0.01
Optimal Parameter:	'svc__c': [0.01, 0.1, 1, 10, 100], 'svc__gamma': [0.01, 0.1, 1, 10, 100]

Table 4. Support Vector Classifier Hyperparameter

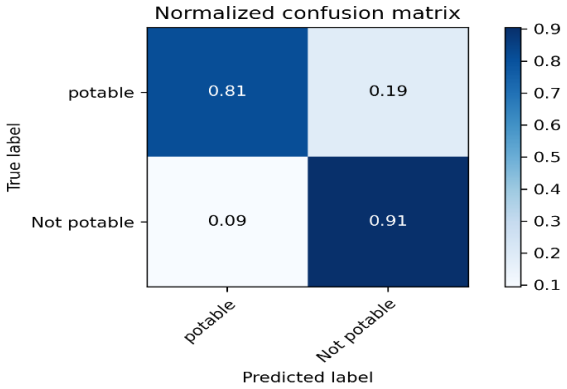

Figure 11. Normalized confusion matrix for Support Vector Classifier

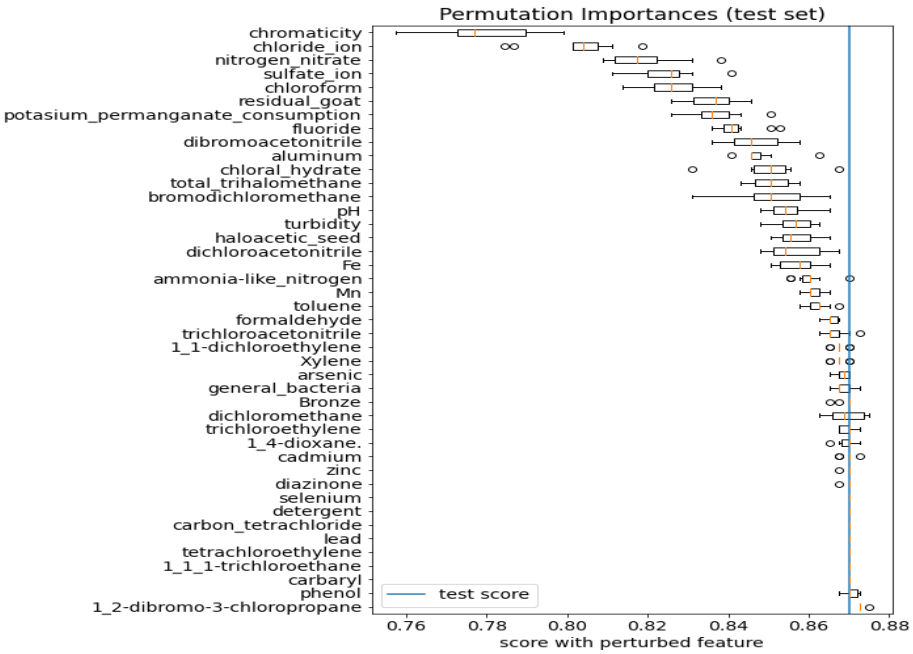

Figure 12. Feature importance for Support Vector Classifier

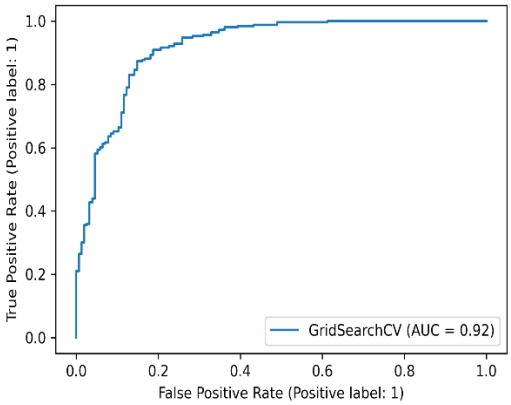

Figure 13. ROC curve for Support Vector Classifier

Using the parameters, above the baseline accuracy of the support vector classifier has been increased by around 2%. Feature importance shows that 1-2 dibromo-3-chloropropane is the most crucial feature. Also, plotting the ROC curve, AUC can reach 0.92.

4-4 XGBoost

Test Baseline:	0.6200980392156863
Test Score:	0.8799019607843137
Standard Deviation of Accuracy:	+/- 0.01
Optimal Parameter:	"xgbclassifier__max_depth" : [1, 3, 5, 10, 30, 50, 100], "xgbclassifier__min_child_weight" : [1, 3, 5, 7], "xgbclassifier__gamma": [0.0, 0.1, 0.2 , 0.3, 0.4] use_label_encoder=False,n_estimators=100, learning_rate=0.2,subsample=0.66, nthread=1

Table 5. XGBoost Hyperparameter

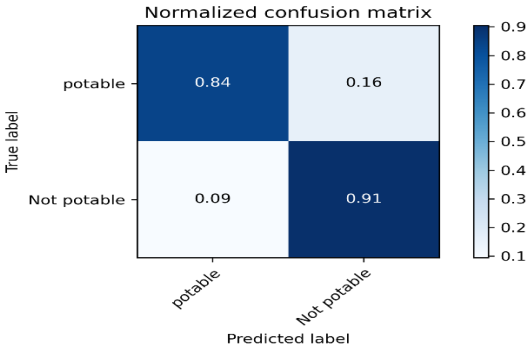

Figure 14. Normalized confusion matrix for XGBoost

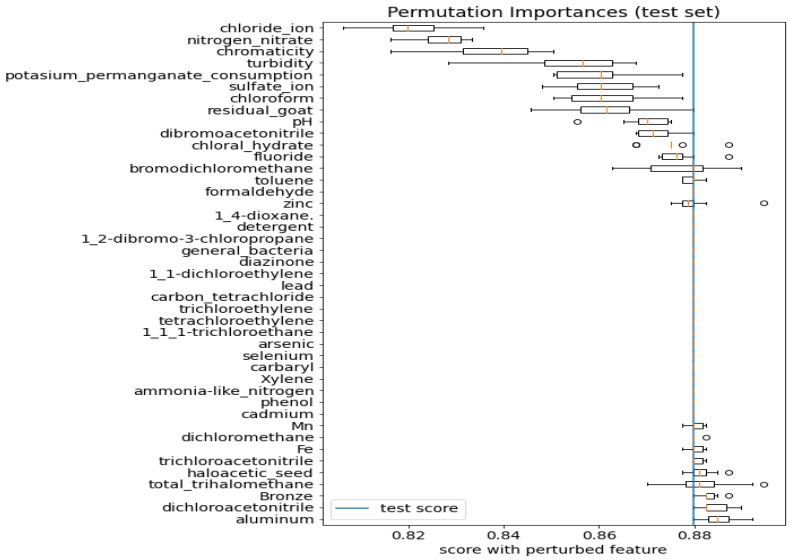

Figure 15. Feature importance for XGBoost

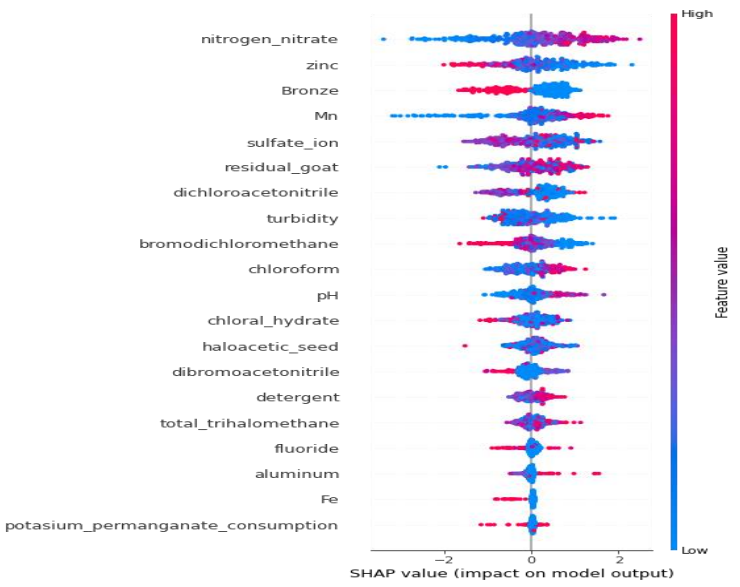

Figure 16. Summary plot for XGBoost using SHAP

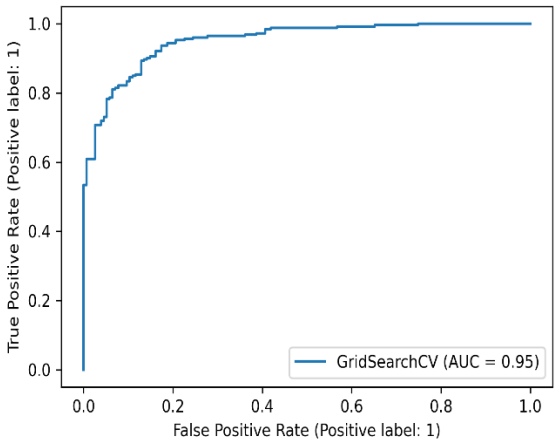

Figure 17. ROC curve for XGBoost Classifier

Using the parameters, above the baseline accuracy of the XGBoost classifier has been increased by around 2%. Feature importance shows that nitrogen nitrate is the most crucial feature using SHAP for global permutation importance aluminum. Also, plotting the ROC curve, it can reach 0.95.

Overall Support vector classifier, Random Forest classifier, and XGBoost classifier can be ideal models for this data set. However, choosing one XGBoost will be the suitable model since it displays the highest AUC, 0.95.

5. Outlook

After examing the research for this dataset, the research came out unsatisfactory and created confusion. Before I got the result, I expected turbidity is the critical factor for feature importance; for example, every model shows different importance results globally and locally. For the tunning parameter, further research is necessary to expand the hyperparameter tunning will likely get a better prediction model. Also, advanced deep-learning methods would likely improve the results to get further improvement. Additionally, I did research using semi-automated machine learning and automated machine learning from outside the rubric spectrum. Unfortunately, it was not mentioned in this report because, due to lack of experience, it could identically copy the hyperparameter tuning that I used in this project. However, in the end, research, this

project gave the experience and insight into machine learning it would indeed be valuable assets for future work.

6. Reference

A Study on the Turbidity Estimation Model Using Data Mining Techniques in the Water Supply System. (n.d.). Retrieved 2021, from https://pdfs.semanticscholar.org/32b2/0c22003f3bc075a61f44287bde0bdb3aeaa9.pdf.

7. Github repository

https://github.com/jinnoh47/data1030-project/tree/main